Technicolor/INRIA team at the MediaEval 2013 Violent Scenes Detection Task

Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, Patrick Gros

Technicolor/INRIA team at the MediaEval 2013 Violent Scenes Detection Task*

Cédric Penet, Claire-Hélène Demarty
Technicolor/INRIA Rennes & Technicolor
1 ave de Belle Fontaine
35510 Cesson-Sévigné, France
cedric.penet@technicolor.com
claire-helene.demarty@technicolor.com

Guillaume Gravier, Patrick Gros
CNRS/IRISA & INRIA Rennes
Campus de Beaulieu
35042 Rennes, France
guig@irisa.fr
Patrick.Gros@inria.fr

ABSTRACT

This paper presents the work done at Technicolor and INRIA regarding the MediaEval 2013 Violent Scenes Detection task, which aims at detecting violent scenes in movies. We participated in both the objective and the subjective subtasks.

1. INTRODUCTION

The MediaEval 2013 Violent Scenes Detection Task is a continuation of the MediaEval 2012 and 2011 Affect Task and aims at detecting violence in movies. A complete description of the task and datasets may be found in [1]. This paper is a joint effort between Technicolor and INRIA (TEC-INRIA). Compared to what was submitted in 2012, a new and improved system towards the generalisation to different movies is proposed. This system is presented in section 2, while the derived runs are detailed in section 3. The results are discussed in section 4.

2. SYSTEM DESCRIPTION

For this year's benchmark, we have developed an improved multimodal system, compared to what we proposed in 2012. It is described on figure 1. The main idea behind this system is to use the output of concept detectors as input to a violence detector and was inspired by the work of the ARF team [6, 2]. It is composed of 5 main steps.

Segment-based audio concept detectors:
Our audio concept detector is described in [3, 4]. We first model the variability between the different movies using factor analysis, and compensate the audio features directly by removing the modeled variability from the features. Once this has been performed, variable length segments are extracted using a data-driven segmentation. Audio words sequences are then computed using three different types of features, namely MFCC, energy and flatness coefficients, extracted on a uniform Mel filterbank.

A naive contextual Bayesian network (either per feature type or with all features altogether) is then used on top of that to classify each audio segment according to its context, i.e., each sample is represented using both its own words, and the words of the n samples before and the n samples after.

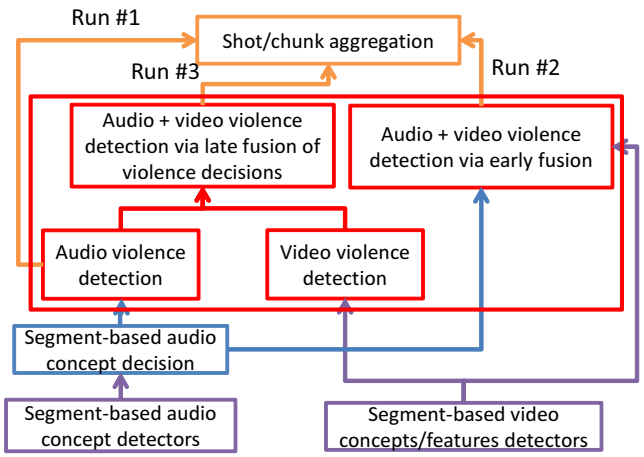

Figure 1: Description of TEC-INRIA multimodal system.

Trained for the detection of screams, gunshots and explosions, our system has proved to be comparable to the state-of-the-art. For each sample s_i, we obtain $P(s_i \in c_k), \forall\, k \in \{\text{gunshots, explosions, screams, others}\}$, where others corresponds to all that is not screams, gunshots or explosions. For more insight, please read [3, 4].

Segment-based audio concepts decision:
Once the probabilities have been estimated, decisions are made using two types of methods. First, a single decision variable is extracted, taking, for each sample, the value of the class that has the highest probability. Second, in order to allow more flexibility in the system, four binary decision variables are extracted, one for each class. The binary variables are set to one if the probability of a sample of belonging to the corresponding class is higher than 30 %. This allows the segments samples to be detected as belonging to several classes at once.

Segment-based video concepts/features detectors:
The video features used in this system are the same as what we used for run #3 of last year's task [5]: shot length, three color harmonisation based features, color coherence, blood-color proportion, flash detection, fire detection, motion intensity, average luminance. These features are all aggregated on the same variable segments as for the audio. For the image based features, aggregation is performed via av-

*We would like to acknowledge the MediaEval Multimedia Benchmark http://www.multimediaeval.org/ and in particular the Violent Scenes Detection Task 2013 for providing the data used in this research. This work was partly achieved as part of the Quaero program, funded by Oseo, french state agency for innovation.

MediaEval 2013 Workshop October 18-19, 2013, Barcelona, Spain

Runs	OBJECTIVE		SUBJECTIVE	
	Shots	Chunks	Shots	Chunks
Run #1	**33.82 %**	-	**53.59 %**	**44.79 %**
Run #2	12.02 %	-	34.00 %	-
Run #3	13.17 %	**12.47 %**	30.22 %	18.81 %
Run #4	22.48 %	-	-	-

Table 1: Results obtained for both the objective and the subjective subtasks in terms of MAP@100, the official metric.

eraging of values, while for shot based features (shot length and color coherence), aggregation is performed by replication of the values. After this aggregation process, all the features are quantised on 21 values.

Violence detection and shot/chunk aggregation:
Once both video and audio concepts have been extracted, naive contextual Bayesian networks are used to detect whether samples are violent or not. Once again, two methods have been tested. First, early fusion of the concepts is performed, i.e., a unique violence classifier is trained using both audio and video concepts. Second, we proceeded to late fusion, by means of two separated audio and video classifiers, whose violence decisions are plugged into an additional naive Bayesian network. After classification, shot aggregation or chunk aggregation is performed. Chunk aggregation is performed by grouping contiguous segments for which the decision is the same. Then, for both shots and chunks, their probability of being violent is set to the highest probability of the segments that lie within the shots/chunks.

3. RUNS SUBMITTED

In this section, we present the runs that we submitted for this year's task. The first three runs are based on our new multimodal system, for which several configurations were chosen using cross-validation experimentation for both the objective and the subjective subtasks. The fourth run corresponds to run #3 of our participation in last year's benchmark [5]. In order to evaluate the stability of this previous system, we re-used the exact same audio and video models, without retraining the parameters.

Run #1: Audio only
The first run is audio only. Audio concepts detection is performed using a context window of size $n = 1$ and violence detection using a single decision variable is performed using a context window of size $n = 5$. In addition, we trained different classifiers on each audio features type, resulting in several audio concept detectors. We then performed late fusion of these classifiers for violence detection using optimal weights fusion. Only shot-level runs obtained through shot aggregation have been submitted for the objective subtask, while both shot and chunk aggregation were used for the subjective one.

Run #2: Multimodal early fusion
This run is an early fusion of audio concepts and video concepts. The audio concepts provided are the audio concepts extracted from each type of audio features. They are extracted using a context window of size $n = 5$. Violence detection is then performed using a context window of size $n = 5$. For the objective subtask, four audio binary nodes corresponding to the four audio concepts decisions are used per features type, while only one is used for the subjective subtask. Only shot aggregation have been submitted for both subtasks.

Run #3: Multimodal late fusion
This run is equivalent to run #2, the main difference being that instead of early fusion, late fusion through a naive Bayesian network is used. Shot and chunk aggregation have been submitted for both subtasks.

Run #4: last year's models
This run has only been submitted with shot aggregation for the objective subtask.

4. RESULTS AND DISCUSSION

The obtained results are reported in table 1 for each submitted run and both subtasks. It must be noted first that, compared to last year's results, where our best system achieved about 62 % in term of MAP@100, this year's results are much lower. Our best MAP@100 for the objective definition is 33.82 %. Moreover, our best achievement is obtained with run #1, and the results we obtained for run #2, #3 and #4 are very poor in comparison. This contradicts the results that were previously obtained about multimodality, especially for run #4, which is a reuse of last year's models. We think this is an indication of a flaw in our multimodal protocol. However, this may also indicate that last year's results obtained on a set of only three movies were may be overly optimistic.

It must also be noted that the results obtained for the subjective definition are much higher than for the objective definition, which indicates that the subjective definition might be less variable than the objective one. Another reason for this may be that globally the duration for subjective violence in the ground truth is bigger than the duration for the objective one. It is also interesting that the results obtained at chunk level are slightly lower than the results obtained at shot level, indicating the importance of temporal integration: the more the system is integrated, the better the results are.

Finally, we obtain good results in terms of recall and precision for most of our runs and for the both violence definitions. For the objective definition, apart from run #4 (2012 system), our shot-level runs reach more than 80 % recall and 20 % precision (runs #2 and #3 even reach recall values of 90 % and 87 % respectively). The subjective definition yields equivalent recall rates, and improved precision rates, up to 30 %, for runs at shot level.

5. REFERENCES

[1] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. L. Quang, and Y.-G. Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *MediaEval Workshop*, 2013.

[2] B. Ionescu, J. Schlüter, I. Mironică, and M. Schedl. A Naïve Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies. In *ACM ICMR*, 2013.

[3] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Audio Event Detection in Movies using Multiple Audio Words and Contextual Bayesian Networks. In *CBMI*, June 2013.

[4] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Variability Modelling for Audio Events Detection in Movies. *MTAP - Special Issue on CBMI*, 2013. Submitted to.

[5] C. Penet, C.-H. Demarty, M. Soleymani, G. Gravier, and P. Gros. Technicolor/INRIA/Imperial College London at the MediaEval 2012 Violent Scene Detection Task. In *MediaEval 2012 Workshop*, 2012.

[6] J. Schlüter, B. Ionescu, I. Mironică, and M. Schedl. ARF @ MediaEval 2012: An Uninformed Approach to Violence Detection in Hollywood Movies. In *MediaEval 2012 Workshop*, 2012.